Docket No.: 31869-202101
(PATENT)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:
Tetsuji Nakagawa

Application No.: Not Yet Assigned

Confirmation No.:

Filed: Concurrently Herewith

Art Unit: N/A

For: MORPHOLOGICAL ANALYZER, MORPHOLOGICAL ANALYSIS METHOD, AND MORPHOLOGICAL ANALYSIS PROGRAM

Examiner: Not Yet Assigned

CLAIM FOR PRIORITY AND SUBMISSION OF DOCUMENTS

MS Patent Application
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

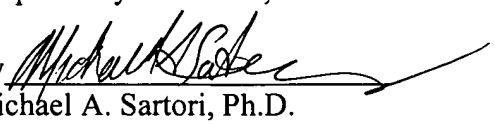
Applicant hereby claims priority under 35 U.S.C. 119 based on the following prior foreign application filed in the following foreign country on the date indicated:

Country	Application No.	Date
Japan	2003-154625	May 30, 2003

In support of this claim, a certified copy of the said original foreign application is filed herewith.

Dated: March 30, 2004

Respectfully submitted,

By 

Michael A. Sartori, Ph.D.
Registration No.: 41,289
VENABLE LLP
P.O. Box 34385
Washington, DC 20043-9998
(202) 344-4000
(202) 344-8300 (Fax)
Attorney/Agent For Applicant

DC2-534294

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年　５月３０日

出願番号
Application Number: 特願２００３－１５４６２５

[ST. 10/C]: [ＪＰ２００３－１５４６２５]

出願人
Applicant(s): 沖電気工業株式会社

２００３年１１月２８日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３０９８５６３

【書類名】　特許願
【整理番号】　KN002592
【提出日】　平成15年 5月30日
【あて先】　特許庁長官　太田　信一郎　殿
【国際特許分類】　G06F 17/27
【発明者】
　【住所又は居所】　東京都港区虎ノ門１丁目７番１２号　沖電気工業株式会社内
　【氏名】　中川　哲治
【特許出願人】
　【識別番号】　000000295
　【氏名又は名称】　沖電気工業株式会社
　【代表者】　篠塚　勝正
【代理人】
　【識別番号】　100090620
　【弁理士】
　【氏名又は名称】　工藤　宣幸
【その他】　国等の委託研究の成果に係る特許出願（平成１４年度通信・放送機構「多言語標準文書処理システムの研究開発」委託研究、産業活力再生特別措置法第３０条の適用を受けるもの）
【手数料の表示】
　【予納台帳番号】　013664
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9006358

【プルーフの要否】　　要

【書類名】　明細書

【発明の名称】　形態素解析装置、形態素解析方法及び形態素解析プログラム

【特許請求の範囲】

【請求項１】　形態素解析対象文に対して所定の形態素解析方法を適用し、活用形がある品詞についてはその活用形の情報を含む品詞タグが付与された単語列でなる、形態素解析結果の候補である仮説を１又は複数生成する仮説生成手段と、

品詞に関する複数種類のｎ－ｇｒａｍ確率モデルの情報を格納しているモデル格納手段と、

上記各仮説に対し、大量の文中でその仮説が出現するであろう生成確率を、上記モデル格納手段に格納されている複数種類のｎ－ｇｒａｍ確率モデルの情報を重み付けて結合して求める生成確率計算手段と、

上記各仮説の生成確率に基づき、解となる仮説を探索する解探索手段とを備え、

上記モデル格納手段は、少なくとも、品詞及び品詞の活用形を反映させた種類のｎ－ｇｒａｍ確率モデルの情報は格納している

ことを特徴とする形態素解析装置。

【請求項２】　品詞及び品詞の活用形を反映させた種類のｎ－ｇｒａｍ確率モデルの情報は、上記仮説のｉ番目の単語ω_iの品詞をt_i、品詞の活用形をt_iｆｏｒｍとしたとき、その品詞t_iをとる単語の中で単語ω_iが出現する条件付き確率Ｐ（$\omega_i \mid t_i$）と、単語ω_iに係る品詞t_iＰＯＳがその活用形t_iｆｏｒｍで出現する条件付き確率Ｐ（t_iｆｏｒｍ｜t_iＰＯＳ）と、直前Ｎ－１個の単語に係る品詞タグ列$t_{i-N+1} \cdots t_{i-1}$の並びに続いて単語ω_iに係る品詞t_iＰＯＳが出現する条件付き確率Ｐ（t_iＰＯＳ｜$t_{i-N+1} \cdots t_{i-1}$）との積になっていることを特徴とする請求項１に記載の形態素解析装置。

【請求項３】　上記モデル格納手段は、複数種類の１種類として、クラスｎ－ｇｒａｍ確率モデルの情報も格納していることを特徴とする請求項１又は２に

記載の形態素解析装置。

【請求項４】　クラスｎ－ｇｒａｍ確率モデルの情報におけるクラスの種別は、品詞タグ付きコーパスと品詞タグ無しコーパスとから学習したものであることを特徴とする請求項３に記載の形態素解析装置。

【請求項５】　形態素解析対象文に対して所定の形態素解析方法を適用し、活用形がある品詞についてはその活用形の情報を含む品詞タグが付与された単語列でなる、形態素解析結果の候補である仮説を１又は複数生成する仮説生成工程と、

上記各仮説に対し、大量の文中でその仮説が出現するであろう生成確率を、予め用意されている、品詞及び品詞の活用形を反映させた種類のｎ－ｇｒａｍ確率モデルの情報を含む、品詞に関する複数種類のｎ－ｇｒａｍ確率モデルの情報を重み付けて結合して求める生成確率計算工程と、

上記各仮説の生成確率に基づき、解となる仮説を探索する解探索工程と

を含むことを特徴とする形態素解析方法。

【請求項６】　請求項５の形態素解析方法を、コンピュータが実行可能なコードで記述していることを特徴とする形態素解析プログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は形態素解析装置、形態素解析方法及び形態素解析プログラムに関し、特に、複数の正解候補の中から最適な解を高い精度で選択し得るようにしたものである。

【０００２】

【従来の技術】

形態素解析装置は、入力された文に対してその文を構成する各形態素を同定して区切り、品詞を付与するものである。しかしながら、形態素に分割する際及び品詞を付与する際に、複数の正解候補が存在し曖昧性が発生するため、正解候補の中から正しいものを選択する必要がある。

【０００３】

このような目的のために、以下のような品詞ｎ－ｇｒａｍモデルに基づく方法がいくつか提案されている。

【０００４】

【特許文献１】　特開平７－２７１７９２号公報

【０００５】

【非特許文献１】　浅原、松本著、形態素解析のための拡張統計モデル」、情処論Ｖｏｌ．４３，Ｎｏ．３，ｐｐ．６８５－６９５，２００２

特許文献１は、日本語形態素解析において、統計的手法によりこの曖昧性を解決する方法について述べている。直前の２つの品詞が与えられたときの３つ目の品詞が出現する確率である品詞三つ組確率と、品詞が与えられたときの単語の出現確率である品詞別単語出力確率から、文を構成する単語列と各単語に付与された品詞列の同時確率を最大にするような候補を選ぶことにより、曖昧性の解消を行っている。

【０００６】

非特許文献１では、特徴的な性質を持つ形態素の品詞を語彙化し、似た性質を持つ品詞をグループ化するという拡張を行うことで、より精度の高い形態素解析を実現している。

【０００７】

【発明が解決しようとする課題】

しかしながら、特許文献１の記載方法は、過去の品詞系列のみから次に来る品詞を予測し、さらに品詞が与えられた場合の条件のみから単語を予測しているため、高い精度で形態素解析を行うのは困難である。つまり、助詞等の機能語はしばしば他の形態素と異なる特徴的な性質をもつが、このような語に関しては品詞だけではなく語彙自体の情報も考慮する必要がある。また、品詞体系によっては数百を越える数の品詞を扱わなければならないこともあるが、そのような場合は品詞の組合わせの数が膨大になるため、特許文献１の記載方法を直接適用して形態素解析を行うことは困難である。

【０００８】

非特許文献１の記載方法では、品詞の語彙化により特徴的な性質を持つ形態素

に対処している。また、品詞のグループ化を行うことにより品詞の数が多い場合にも対処している。しかしながら、語彙化やグループ化は誤り駆動に基づく方法を用いて一部の形態素や品詞に関してのみ行われるため、形態素に関する十分な情報を利用できているわけではなく、また、訓練データを効果的に利用できないという課題がある。

【０００９】

そのため、複数の正解候補の中から最適な解を高い精度で選択し得る形態素解析装置、形態素解析方法及び形態素解析プログラムが望まれている。

【００１０】

【課題を解決するための手段】

かかる課題を解決するため、第１の本発明の形態素解析装置は、（１）形態素解析対象文に対して所定の形態素解析方法を適用し、活用形がある品詞についてはその活用形の情報を含む品詞タグが付与された単語列でなる、形態素解析結果の候補である仮説を１又は複数生成する仮説生成手段と、（２）品詞に関する複数種類のｎ－ｇｒａｍ確率モデルの情報を格納しているモデル格納手段と、（３）上記各仮説に対し、大量の文中でその仮説が出現するであろう生成確率を、上記モデル格納手段に格納されている複数種類のｎ－ｇｒａｍ確率モデルの情報を重み付けて結合して求める生成確率計算手段と、（４）上記各仮説の生成確率に基づき、解となる仮説を探索する解探索手段とを備え、（２－１）上記モデル格納手段が、少なくとも、品詞及び品詞の活用形を反映させた種類のｎ－ｇｒａｍ確率モデルの情報は格納していることを特徴とする。

【００１１】

第２の本発明の形態素解析方法は、（１）形態素解析対象文に対して所定の形態素解析方法を適用し、活用形がある品詞についてはその活用形の情報を含む品詞タグが付与された単語列でなる、形態素解析結果の候補である仮説を１又は複数生成する仮説生成工程と、（２）上記各仮説に対し、大量の文中でその仮説が出現するであろう生成確率を、予め用意されている、品詞及び品詞の活用形を反映させた種類のｎ－ｇｒａｍ確率モデルの情報を含む、品詞に関する複数種類のｎ－ｇｒａｍ確率モデルの情報を重み付けて結合して求める生成確率計算工程と

、（３）上記各仮説の生成確率に基づき、解となる仮説を探索する解探索工程とを含むことを特徴とする。

【００１２】

第２の本発明の形態素解析プログラムは、第２の本発明の形態素解析方法を、コンピュータが実行可能なコードで記述していることを特徴とする。

【００１３】

【発明の実施の形態】

（Ａ）第１の実施形態

以下、本発明による形態素解析装置、形態素解析方法及び形態素解析プログラムの第１の実施形態を図面を参照しながら説明する。

【００１４】

（Ａ－１）第１の実施形態の構成

図１は、第１の実施形態の形態素解析装置の機能的構成を示すブロック図である。第１の実施形態の形態素解析装置は、例えば、入出力装置や補助記憶装置などを備えるパソコン等の情報処理装置上に、形態素解析プログラム（図２～図４参照）をインストールすることによって実現されるが、機能的には、図１で表すことができる。

【００１５】

第１の実施形態の形態素解析装置１００は、大きくは、確率モデルを使用して形態素解析を行う解析部１１０、確率モデル等を格納するモデル格納部１２０、及び、パラメータ学習用のコーパスから確率的モデルの学習を行うためのモデル学習部１３０から構成されている。

【００１６】

解析部１１０は、形態素解析を行う文を入力するための入力部１１１、入力された文に対して、形態素辞書格納部１２１に格納されている形態素辞書を用いて可能な解（形態素解析結果）の候補（仮説）を生成する仮説生成部１１２、生成された各仮説に対して、確率モデル格納部１２２に格納された品詞ｎ－ｇｒａｍモデル、語彙化品詞ｎ－ｇｒａｍモデル（当該モデルの定義については後述する）及び階層化品詞ｎ－ｇｒａｍモデル（当該モデルの定義については後述する）

を、重み格納部１２３に格納された重み付けにより結合して生成確率を計算する生成確率計算部１１３、生成確率の付与された仮説の中から最も尤度の高い解を選ぶ解探索部１１４、及び、解探索部１１４により得られた解を出力する出力部１１５より構成される。

【００１７】

なお、入力部１１１は、例えば、キーボード等の一般的な入力部だけでなく、記録媒体のアクセス装置等のファイル読込装置や、文書をイメージデータとして読み込んでそれをテキストデータに置き換える文字認識装置等も該当する。また、出力部１１５は、例えば、ディスプレイやプリンタ等の一般的な出力部だけでなく、記録媒体へ格納する記録媒体アクセス装置等も該当する。

【００１８】

モデル格納部１２０は、確率推定部１３２で計算され、生成確率計算部１１３及び重み計算部１３３で使用される確率モデルを格納した確率モデル格納部１２２、重み計算部１３３で計算され、生成確率計算部１１３で使用される重みを格納する重み格納部１２３、及び、仮説生成部１１２で解候補（仮説）を生成するために使用される形態素辞書を格納する形態素辞書格納部１２１から構成されている。

【００１９】

モデル学習部１３０は、確率推定部１３２及び重み計算部１３３でモデルの学習を行うために使用される品詞タグ付きコーパス格納部１３１、品詞タグ付きコーパス格納部１３１に格納された品詞タグ付きコーパスを用いて確率モデルの推定を行い、その結果を確率モデル格納部１２２へ格納する確率推定部１３２、及び、確率モデル格納部１２２に格納された確率モデルと品詞タグ付きコーパス格納部１３１に格納された品詞タグ付きコーパスを用いて確率モデルの重みを計算し、その結果を重み格納部１２３へ格納する重み計算部１３３から構成されている。

【００２０】

（Ａ－２）第１の実施形態の動作

次に、第１の実施形態の形態素解析装置１００の動作（第１の実施形態の形態

素解析方法）を、図２のフローチャートを参照しながら説明する。図２は、入力された文を形態素解析装置１００が形態素解析して出力するまでの処理の流れを示すフローチャートである。

【００２１】

まず、使用者が入力した形態素解析をしたい文を入力部１１１によって取り込む（２０１）。入力された文に対して、仮説生成部１１２は、形態素辞書格納部１２１に格納された形態素辞書を用いて、可能な解の候補である仮説を生成する（２０２）。この仮説生成部１１２による処理は、例えば、一般的な形態素解析方法を適用する。生成確率計算部１１３は、確率モデル格納部１２２及び重み格納部１２３に格納された情報を用いて、仮説生成部１１２で生成された各仮説に対しその生成確率を計算する（２０３）。生成確率計算部１１３は、各仮説に対する生成確率として、品詞ｎ－ｇｒａｍ、語彙化品詞ｎ－ｇｒａｍ及び階層化品詞ｎ－ｇｒａｍを確率的に重み付けたものを計算する。

【００２２】

ここで、入力された文の先頭から（ｉ＋１）番目の単語及びその品詞タグをそれぞれωｉ及びｔｉとし、文中の単語（形態素）の数をｎとする。また、品詞タグｔは、品詞ｔＰＯＳと活用形ｔｆｏｒｍからなっているとする。なお、活用形がない品詞の場合には、品詞と品詞タグとは同一のものである。仮説、つまり正解候補の単語・品詞タグ列は、

ω０ｔ０　…　ωｎ－１ｔｎ－１

と表現され、このような仮説の中から最も生成確率の高いものを解として選べばよいため、（１）式を満足する最適な単語・品詞タグ列を求めることになる。

【００２３】

例えば、「私は見た。」という文章は、「私（名詞；より細かく分類した代名詞を適用しても良い）／は（助詞；より細かく分類した副助詞を適用しても良い）／見（動詞－連用形）／た（助動詞）／。（句点）」という単語・品詞タグ列と、「私（名詞）／は（助詞）／見（動詞－終止形）／た（助動詞）／。（句点）」という単語・品詞タグ列との２つの仮説が生じ、いずれが最適であるかが（１）式によって求められる。なお、この例の場合、「見」に関してのみ、「動詞

」という品詞と「連用形」又は「終止形」という活用形で品詞タグが構成され、他の単語（句点も１個の単語として取扱う）については品詞のみで品詞タグが構成されている。

【００２４】

【数１】

$$
\begin{aligned}
&\hat{w}_0\hat{t}_0\cdots\hat{w}_{n-1}\hat{t}_{n-1}\\
&= \underset{w_0t_0\cdots w_{n-1}t_{n-1}}{\operatorname{argmax}} P(w_0t_0\cdots w_{n-1}t_{n-1})\\
&= \underset{w_0t_0\cdots w_{n-1}t_{n-1}}{\operatorname{argmax}} \prod_{i=0}^{n-1} P(w_it_i|w_0t_0\cdots w_{i-1}t_{i-1})\\
&= \underset{w_0t_0\cdots w_{n-1}t_{n-1}}{\operatorname{argmax}} \prod_{i=0}^{n-1} \sum_{\mathcal{M}\in\mathbf{M}} P(\mathcal{M}|w_0t_0\cdots w_{i-1}t_{i-1})P(w_it_i|w_0t_0\cdots w_{i-1}t_{i-1}\mathcal{M}) \qquad (1)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{M} = \{&\mathcal{M}^1_{\mathrm{POS}},\cdots,\mathcal{M}^{N_{\mathrm{POS}}}_{\mathrm{POS}},\\
&\mathcal{M}^1_{\mathrm{lex1}},\cdots,\mathcal{M}^{N_{\mathrm{lex1}}}_{\mathrm{lex1}},\mathcal{M}^1_{\mathrm{lex2}},\cdots,\mathcal{M}^{N_{\mathrm{lex2}}}_{\mathrm{lex2}},\mathcal{M}^1_{\mathrm{lex3}},\cdots,\mathcal{M}^{N_{\mathrm{lex3}}}_{\mathrm{lex3}},\\
&\mathcal{M}^1_{\mathrm{hier}},\cdots,\mathcal{M}^{N_{\mathrm{hier}}}_{\mathrm{hier}}\} \qquad (2)
\end{aligned}
$$

$$
\sum_{\mathcal{M}\in\mathbf{M}} P(\mathcal{M}) = 1 \qquad (2.5)
$$

（１）式において、第１行の「＾ω０＾ｔ０　…　＾ωｎ－１＾ｔｎ－１」は最適な単語・品詞タグ列を意味しており、ａｒｇｍａｘは、複数の単語・品詞タグ列（仮説）の中から生成確率Ｐ（ω０ｔ０　…　ωｎ－１ｔｎ－１）が最も高い単語・品詞タグ列を選択することを表している。

【００２５】

ある単語・品詞タグ列の生成確率Ｐ（ω０ｔ０　…　ωｎ－１ｔｎ－１）は、その単語・品詞タグ列においてその（ｉ＋１）番目（ｉは０～（ｎ－１））の単語・品詞タグが生じる条件付き確率Ｐ（ωｉｔｉ｜ω０ｔ０　…　ωｎ－１ｔｎ－１）の積で表される。条件付き確率Ｐ（ωｉｔｉ｜ω０ｔ０　…　ωｎ－１ｔｎ－１）は、あるｎ－ｇｒａｍモデルＭで計算される単語についての出力確率Ｐ（ωｉｔｉ｜ω０ｔ０　…　ωｎ－１ｔｎ－１Ｍ）と、そのｎ－ｇｒａｍモデルＭに対する重みＰ（Ｍ｜ω０ｔ０　…　ωｎ－１ｔｎ－１）との積を、全てのモデルについて求めた積和で表される。

【００２６】

ここで、出力確率Ｐ（$\omega_i t_i | \omega_0 t_0 \cdots \omega_{n-1} t_{n-1} M$）の情報が、確率モデル格納部１２２に格納されており、ｎ－ｇｒａｍモデルＭに対する重みＰ（$M | \omega_0 t_0 \cdots \omega_{n-1} t_{n-1}$）の情報が、重み格納部１２３に格納されていいる。

【００２７】

（２）式は、生成確率Ｐ（$\omega_0 t_0 \cdots \omega_{n-1} t_{n-1}$）の計算に適用される全てのモデルＭを集合Ｍとして記載したものである。但し、集合Ｍは、（２．５）式に示すように、その要素である各モデルＭ毎の確率Ｐ（Ｍ）が１になるようなモデルの集合である。

【００２８】

モデルＭについての下付パラメータはモデルの種類を表しており、「ＰＯＳ」は品詞ｎ－ｇｒａｍモデルを表しており、「ｌｅｘ１」は第１の語彙化品詞ｎ－ｇｒａｍモデルを表しており、「ｌｅｘ２」は第２の語彙化品詞ｎ－ｇｒａｍモデルを表しており、「ｌｅｘ３」は第３の語彙化品詞ｎ－ｇｒａｍモデルを表しており、「ｈｉｅｒ」は階層化品詞ｎ－ｇｒａｍモデルを表している。モデルＭについての上付パラメータは、そのモデルにおける記憶長の長さＮ－１、言い換えると、ｎ－ｇｒａｍでの単語数（品詞タグ数も同数）を表している。

【００２９】

【数２】

$\mathcal{M}^N_{\mathrm{POS}}$: **品詞 N-gram モデル**

$$P(w_it_i|w_0t_0\cdots w_{i-1}t_{i-1}\mathcal{M}^N_{\mathrm{POS}}) \equiv P(w_i|t_i)P(t_i|t_{i-N+1}\cdots t_{i-1}) \quad (3)$$

$\mathcal{M}^N_{\mathrm{lex1}}, \mathcal{M}^N_{\mathrm{lex2}}, \mathcal{M}^N_{\mathrm{lex3}}$: **語彙化品詞 N-gram モデル**

$$P(w_it_i|w_0t_0\cdots w_{i-1}t_{i-1}\mathcal{M}^N_{\mathrm{lex1}}) \equiv P(w_i|t_i)P(t_i|w_{i-N+1}t_{i-N+1}\cdots w_{i-1}t_{i-1}) \quad (4)$$

$$P(w_it_i|w_0t_0\cdots w_{i-1}t_{i-1}\mathcal{M}^N_{\mathrm{lex2}}) \equiv P(w_it_i|t_{i-N+1}\cdots t_{i-1}) \quad (5)$$

$$P(w_it_i|w_0t_0\cdots w_{i-1}t_{i-1}\mathcal{M}^N_{\mathrm{lex3}}) \equiv P(w_it_i|w_{i-N+1}t_{i-N+1}\cdots w_{i-1}t_{i-1}) \quad (6)$$

$\mathcal{M}^N_{\mathrm{hier}}$: **階層化品詞 N-gram モデル**

$$P(w_it_i|w_0t_0\cdots w_{i-1}t_{i-1}\mathcal{M}^N_{\mathrm{hier}}) \equiv P(w_i|t_i)P(t_i^{\mathrm{form}}|t_i^{\mathrm{POS}})P(t_i^{\mathrm{POS}}|t_{i-N+1}\cdots t_{i-1}) \quad (7)$$

記憶長の長さＮ－１の品詞ｎ－ｇｒａｍモデルは、（３）式で定義される。記憶長の長さＮ－１の品詞ｎ－ｇｒａｍモデルは、品詞タグｔｉをとる中でその単語ωｉが出現する条件付き確率Ｐ（ωｉ｜ｔｉ）と、直前Ｎ－１個の単語に係る品詞タグ列ｔｉ－Ｎ＋１…ｔｉ－１の並びに続いてその単語ωｉの品詞タグｔｉが出現する条件付き確率Ｐ（ｔｉ｜ｔｉ－Ｎ＋１…ｔｉ－１）との積で定義される。

【００３０】

記憶長の長さＮ－１の第１の語彙化品詞ｎ－ｇｒａｍモデルは、（４）式で定義される。記憶長の長さＮ－１の第１の語彙化品詞ｎ－ｇｒａｍモデルは、品詞タグｔｉをとる中でその単語ωｉが出現する条件付き確率Ｐ（ωｉ｜ｔｉ）と、直前Ｎ－１個の単語・品詞タグ列ωｉ－Ｎ＋１ｔｉ－Ｎ＋１…ωｉ－１ｔｉ－１の並びに続いて、その単語ωｉの品詞タグｔｉが出現する条件付き確率Ｐ（ｔｉ｜ωｉ－Ｎ＋１ｔｉ－Ｎ＋１…ωｉ－１ｔｉ－１）との積で定義される。

【００３１】

記憶長の長さＮ－１の第２の語彙化品詞ｎ－ｇｒａｍモデルは、（５）式で定義される。記憶長の長さＮ－１の第２の語彙化品詞ｎ－ｇｒａｍモデルは、直前Ｎ－１個の単語に係る品詞タグ列ｔｉ－Ｎ＋１…ｔｉ－１の並びに続いて、単語ωｉとその品詞タグｔｉとの組み合わせωｉｔｉが出現する条件付き確率Ｐ（ω

$_{i}t_{i}|t_{i-N+1}\cdots t_{i-1}$）で定義される。

【００３２】

記憶長の長さＮ－１の第３の語彙化品詞ｎ－ｇｒａｍモデルは、（６）式で定義される。記憶長の長さＮ－１の第３の語彙化品詞ｎ－ｇｒａｍモデルは、直前Ｎ－１個の単語・品詞タグ列$\omega_{i-N+1}t_{i-N+1}\cdots\omega_{i-1}t_{i-1}$の並びに続いて、単語$\omega_{i}$とその品詞タグ$t_{i}$との組み合わせ$\omega_{i}t_{i}$が出現する条件付き確率Ｐ（$\omega_{i}t_{i}|\omega_{i-N+1}t_{i-N+1}\cdots\omega_{i-1}t_{i-1}$）で定義される。

【００３３】

記憶長の長さＮ－１の階層化品詞ｎ－ｇｒａｍモデルは、（７）式で定義される。記憶長の長さＮ－１の階層化品詞ｎ－ｇｒａｍモデルは、その品詞t_{i}をとる単語の中で候補単語ω_{i}が出現する条件付き確率Ｐ（$\omega_{i}|t_{i}$）と、単語ω_{i}に係る品詞t_{i}ＰＯＳがその活用形t_{i}ｆｏｒｍで出現する条件付き確率Ｐ（t_{i}ｆｏｒｍ｜t_{i}ＰＯＳ）と、直前Ｎ－１個の単語に係る品詞タグ列$t_{i-N+1}\cdots t_{i-1}$の並びに続いて単語ω_{i}に係る品詞t_{i}ＰＯＳが出現する条件付き確率Ｐ（t_{i}ＰＯＳ｜$t_{i-N+1}\cdots t_{i-1}$）との積で定義される。なお、単語ω_{i}に係る品詞t_{i}ＰＯＳがその活用形t_{i}ｆｏｒｍで出現する条件付き確率Ｐ（t_{i}ｆｏｒｍ｜t_{i}ＰＯＳ）は、活用形が存在しない品詞については常に「１」として取扱う。

【００３４】

生成確率計算部１１３によって、各仮説に対する生成確率Ｐ（$\omega_{0}t_{0}$ … $\omega_{n-1}t_{n-1}$）が計算されると、解探索部１１４は、（１）式に示すように、その中で最も生成確率が高い解を選択する（図２の２０４）。

【００３５】

上述したように、生成確率計算部１１３による、各仮説に対する生成確率Ｐ（$\omega_{0}t_{0}$ … $\omega_{n-1}t_{n-1}$）の計算を行った後に、解探索部１１４による最も生成確率が高い解（最適解）の探索を行っても良いが、例えば、ビタビ（Ｖｉｔｅｒｂｉ）アルゴリズムを適用して、生成確率計算部１１３による処理と、解探索部１１４による処理とを融合して行うようにしても良い。すなわち、入力

された文の先頭から（ｉ＋１）番目までの単語・品詞タグ列を規定するパラメータｉを徐々に大きくしながら行う、ビタビアルゴリズムによる最適な単語・品詞タグ列の探索によって、生成確率計算部１１３による処理と、解探索部１１４による処理とを融合して行って、最適解を探索する。

【００３６】

上述した（１）式を満足する最適解の単語・品詞タグ列が求まると、出力部１１５によって、求まった最適解（形態素解析結果）をユーザへ出力する（２０５）。

【００３７】

次に、モデル学習部１３０の動作、すなわち、生成確率計算部１１３において使用する確率モデル及び確率モデルの重みを、予め用意された品詞タグ付きコーパスから計算して求める動作を、図３を参照しながら説明する。

【００３８】

まず、確率推定部１３２により、以下に示す確率モデルのパラメータを学習する（３０１）。

【００３９】

ここで、単語列、品詞列、品詞タグ列、及び又は、単語・品詞タグ列などの系列をＸとし、その系列Ｘが品詞タグ付きコーパス格納部１３１に格納されたコーパス中に出現した回数をｆ（Ｘ）で表すと、各確率モデルに対するパラメータは、以下のように表される。

【００４０】

【数３】

$\mathcal{M}^N_{\mathrm{POS}}$: **品詞 N-gram モデル**

$$P(w_i|t_i) = \frac{f(t_i w_i)}{f(t_i)} \quad (8)$$

$$P(t_i|t_{i-N+1}\cdots t_{i-1}) = \frac{f(t_{i-N+1}\cdots t_{i-1}t_i)}{f(t_{i-N+1}\cdots t_{i-1})} \quad (9)$$

$\mathcal{M}^N_{\mathrm{lex1}}, \mathcal{M}^N_{\mathrm{lex2}}, \mathcal{M}^N_{\mathrm{lex3}}$: **語彙化品詞 N-gram モデル**

$$P(w_i|t_i) = \frac{f(t_i w_i)}{f(t_i)} \quad (10)$$

$$P(t_i|w_{i-N+1}t_{i-N+1}\cdots w_{i-1}t_{i-1}) = \frac{f(w_{i-N+1}t_{i-N+1}\cdots w_{i-1}t_{i-1}t_i)}{f(w_{i-N+1}t_{i-N+1}\cdots w_{i-1}t_{i-1})} \quad (11)$$

$$P(w_i t_i|t_{i-N+1}\cdots t_{i-1}) = \frac{f(t_{i-N+1}\cdots t_{i-1}w_i t_i)}{f(t_{i-N+1}\cdots t_{i-1})} \quad (12)$$

$$P(w_i t_i|w_{i-N+1}t_{i-N+1}\cdots w_{i-1}t_{i-1}) = \frac{f(w_{i-N+1}t_{i-N+1}\cdots w_{i-1}t_{i-1}w_i t_i)}{f(w_{i-N+1}t_{i-N+1}\cdots w_{i-1}t_{i-1})} \quad (13)$$

$\mathcal{M}^N_{\mathrm{hier}}$: **階層化品詞 N-gram モデル**

$$P(w_i|t_i) = \frac{f(t_i^{\mathrm{POS}}w_i)}{f(t_i^{\mathrm{POS}})} \quad (14)$$

$$P(t_i^{\mathrm{form}}|t_i^{\mathrm{POS}}) = \frac{f(t_i^{\mathrm{POS}}t_i^{\mathrm{form}})}{f(t_i^{\mathrm{POS}})} \quad (15)$$

$$P(t_i^{\mathrm{POS}}|t_{i-N+1}\cdots t_{i-1}) = \frac{f(t_{i-N+1}\cdots t_{i-1}t_i^{\mathrm{POS}})}{f(t_{i-N+1}\cdots t_{i-1})} \quad (16)$$

記憶長の長さＮ－１の品詞ｎ－ｇｒａｍモデルは、上述したように、（３）式で表されるので、（３）式の右辺の各要素Ｐ（ω ｉ｜ｔ ｉ）及びＰ（ｔ ｉ｜ｔ ｉ－Ｎ＋１…ｔ ｉ－１）を、（８）式及び（９）式に従ってパラメータとして得る。

【００４１】

また、記憶長の長さＮ－１の第１～第３の語彙化品詞ｎ－ｇｒａｍモデルは、上述したように、（４）式～（６）式で表されるので、（４）式～（６）式の右辺の各要素Ｐ（ω ｉ｜ｔ ｉ）、Ｐ（ｔ ｉ｜ω ｉ－Ｎ＋１ｔ ｉ－Ｎ＋１…ω ｉ－１ｔ ｉ－１）、Ｐ（ω ｉｔ ｉ｜ｔ ｉ－Ｎ＋１…ｔ ｉ－１）及びＰ（ω ｉｔ ｉ｜ω ｉ－Ｎ＋１ｔ ｉ－Ｎ＋１…ω ｉ－１ｔ ｉ－１）を、（１０）式～（１３）式に従ってパラメータとして得る。

【００４２】

さらに、記憶長の長さＮ－１の階層化品詞ｎ－ｇｒａｍモデルは、上述したように、（７）式で表されるので、（７）式の右辺の各要素Ｐ（$\omega_i | t_i$）、Ｐ（t_i ｆｏｒｍ｜t_i ＰＯＳ）及びＰ（t_i ＰＯＳ｜$t_{i-N+1} \cdots t_{i-1}$）を、（１４）式～（１６）式に従ってパラメータとして得る。

【００４３】

いずれのパラメータも、コーパス中に、該当する単語列、品詞列、品詞タグ列などが出現した回数を数え上げ、その出現回数、及び又は、各式の分子となる出現回数を分母となる出現回数で除算した値を確率モデル格納部１２２へ格納する。

【００４４】

図５～図７は、確率モデル格納部１２２に格納された一部の確率モデルのパラメータを示す図面である。

【００４５】

次に、品詞タグ付きコーパス格納部１３１に格納されている品詞タグ付きコーパスと確率モデル格納部１２２に格納された確率モデルを用いて、重み計算部１３３により、各確率モデルに対する重みの計算を行い、その結果を重み格納部１２３へ格納する（３０２；図４参照）。

【００４６】

ここで、重みの計算については、（１７）式に示すように、単語・品詞タグ列に依存しない近似を行うこととする。そして、ｌｅａｖｅ－ｏｎｅ－ｏｕｔ法に基づいて、図４に示す手順で計算を行う。

【００４７】

【数４】

$$P(\mathcal{M}|w_0t_0 \cdots w_{i-1}t_{i-1}) \simeq P(\mathcal{M}) \qquad (17)$$

まずはじめに、各モデルＭに対する重みパラメータλ（Ｍ）を全て０にする初期化を行う（４０１）。次に、品詞タグ付きコーパス格納部１３１に格納されている品詞タグ付きコーパスから、単語と品詞タグの対を１つ取り出してω_0 ｔ０

とし、そのｉ個前にある単語と品詞をそれぞれω－１ｔ－１とする（４０２）。次に、各確率モデルＭに対して確率Ｐ’（ω０ｔ０｜ω－Ｎ＋１ｔ－Ｎ＋１…ω－１ｔ－１Ｍ）を計算する（４０３）。

【００４８】

ここで、確率Ｐ’（Ｘ｜Ｙ）＝Ｐ’（ω０ｔ０｜ω－Ｎ＋１ｔ－Ｎ＋１…ω－１ｔ－１Ｍ）は、現在考慮している事象を数え上げの対象から除いて求めた確率値で、（１８）式のようにコーパス中に出現した事象の数を用いて計算する。

【００４９】

【数５】

$$P'(X|Y)=\begin{cases}0 & (f(Y)-1=0)\\ \frac{f(XY)-1}{f(Y)-1} & \text{otherwise.}\end{cases} \qquad (18)$$

以上のようにして各モデルに対し計算した確率値の中で、最も高い値を返したモデルをＭ’とすると、このモデルに対する重みパラメータλ（Ｍ’）を１だけ増やす（４０４）。ステップ４０２～４０４でなる処理を、品詞タグ付きコーパス中の全ての単語と品詞タグとの対について繰り返し（４０５）、全ての単語と品詞タグとの対に対する処理が終了すると、各確率モデルＭに対して、（１９）式に示す正規化した重みＰ（Ｍ）を求める（４０６）。

【００５０】

【数６】

$$P(\mathcal{M})=\frac{\lambda(\mathcal{M})}{\sum_N \lambda(N)} \qquad (19)$$

なお、上記では、簡単のために、（１７）式のように重みの計算に近似を用いたが、かわりに品詞ｎ－ｇｒａｍ、語彙化ｎ－ｇｒａｍ及び階層化品詞ｎ－ｇｒａｍ等の結合を用いて、（１）式と同様に重みを計算することもできる。

【００５１】

（Ａ－３）第１の実施形態の効果

上記第１の実施形態によれば、形態素辞書を利用して得た複数の形態素解析結

果（仮説）から最尤のものを決定する際に、品詞の情報に加え、品詞を語彙化した情報、及び、品詞の階層を考慮した情報を使用してその仮説の生成確率を計算して最尤なものを決定するようにしたので、品詞の情報のみを使用して生成確率を計算して最尤な仮説を決定する方法に比べ、より頑健で高精度な解析を行うことができ、曖昧性を解消できる。

【００５２】

（Ｂ）第２の実施形態

次に、本発明による形態素解析装置、形態素解析方法及び形態素解析プログラムの第１の実施形態を図面を参照しながら説明する。

【００５３】

（Ｂ－１）第２の実施形態の構成

図８は、第２の実施形態の形態素解析装置の機能的構成を示すブロック図である。第２の実施形態の形態素解析装置も、例えば、入出力装置や補助記憶装置などを備えるパソコン等の情報処理装置上に、形態素解析プログラム（図９～図１１参照）をインストールすることによって実現されるが、機能的には、図８で表すことができる。

【００５４】

第２の実施形態の形態素解析装置５００は、大きく見た場合には、第１の実施形態の構成にクラスタリング部５４０が加わったものであり、また、モデル学習部５３０においても、第１の実施形態の構成に、品詞タグ無しコーパス格納部５３４及び品詞タグ・クラス付きコーパス格納部５３５が加わったものである。

【００５５】

クラスタリング部５４０は、クラス学習部５４１、クラスタリングパラメータ格納部５４２及びクラス付与部５４３を有する。

【００５６】

クラス学習部５４１は、品詞タグ付きコーパス格納部５３１中に格納されている品詞タグ付きコーパス及び品詞タグ無しコーパス格納部５３４に格納されている品詞タグ無しコーパスを用いてクラスの学習を行い、学習の結果得られたクラスタリング用のパラメータをクラスタリングパラメータ格納部５４２へ格納する

ものである。

【００５７】

クラス付与部５４３は、クラスタリングパラメータ格納部５４２に格納されているクラスタリング用のパラメータを用いて、品詞タグ付きコーパス格納部５３１中の品詞タグ付きコーパスを入力し、これにクラスを付与したものを品詞タグ・クラス付きコーパス格納部５３５へ格納し、また、仮説生成部５１２で得られた仮説を入力し、これにクラスを付与したものを生成確率計算部５１３へ出力するものである。

【００５８】

品詞タグ・クラス付きコーパス格納部５３５に格納された品詞タグ・クラス付きコーパスは、確率推定部５３２及び重み計算部５３３が利用する。

【００５９】

（Ｂ－２）第２の実施形態の動作

次に、第２の実施形態の形態素解析装置５００の動作（第２の実施形態の形態素解析方法）を、図９のフローチャートを参照しながら説明する。図９は、入力された文を形態素解析装置５００が形態素解析して出力するまでの処理の流れを示すフローチャートである。

【００６０】

第２の実施形態の形態素解析装置５００は、第１の実施形態と比べて、確率値の計算にクラス情報を用いる点だけが異なるため、以下では、第１の実施形態と異なる点についてのみ説明する。

【００６１】

文の入力（６０１）、仮説の生成（６０２）が行われた後、生成された仮説をクラス付与部５４３へ入力してクラスの付与を行い、そのクラスが付与された仮説が生成確率計算部５１３に与えられる（６０３）。クラスの付与の方法については後述する。

【００６２】

次に、クラスが付与された各仮説に対して、生成確率計算部５１３で生成確率の計算を行う（６０４）。但し、各仮説に対する生成確率は、品詞ｎ－ｇｒａｍ

、語彙化品詞ｎ－ｇｒａｍ、階層化品詞ｎ－ｇｒａｍ及びクラス品詞ｎ－ｇｒａｍを確率的に重み付けたものを用いる。計算方法は、上述した（１）式で表されるが、モデルの集合Ｍとして、（２）式に代え、次の（２０）式に示すものが適用される。但し、集合Ｍは、（２０．５）式に示すように、その要素である各モデルＭ毎の確率Ｐ（Ｍ）が１になるようなモデルの集合である。

【００６３】

【数７】

$$\begin{aligned}\mathbf{M} = \{&\mathcal{M}^1_{\mathrm{POS}},\cdots,\mathcal{M}^{N_{\mathrm{POS}}}_{\mathrm{POS}},\\ &\mathcal{M}^1_{\mathrm{lex1}},\cdots,\mathcal{M}^{N_{\mathrm{lex1}}}_{\mathrm{lex1}},\mathcal{M}^1_{\mathrm{lex2}},\cdots,\mathcal{M}^{N_{\mathrm{lex2}}}_{\mathrm{lex2}},\mathcal{M}^1_{\mathrm{lex3}},\cdots,\mathcal{M}^{N_{\mathrm{lex3}}}_{\mathrm{lex3}},\\ &\mathcal{M}^1_{\mathrm{hier}},\cdots,\mathcal{M}^{N_{\mathrm{hier}}}_{\mathrm{hier}},\\ &\mathcal{M}^1_{\mathrm{class1}},\cdots,\mathcal{M}^{N_{\mathrm{class1}}}_{\mathrm{class1}},\mathcal{M}^1_{\mathrm{class2}},\cdots,\mathcal{M}^{N_{\mathrm{class2}}}_{\mathrm{class2}}\}\end{aligned} \tag{20}$$

$$\sum_{\mathcal{M}\in\mathbf{M}} P(\mathcal{M})=1 \tag{20.5}$$

（２）式及び（２０）式の比較から明らかなように、第２の実施形態においては、第１及び第２のクラス品詞ｎ－ｇｒａｍモデルも適用されている。

【００６４】

（２０）式において、下付パラメータが「ｃｌａｓｓ１」のものが第１のクラス品詞ｎ－ｇｒａｍモデルを表しており、下付パラメータが「ｃｌａｓｓ２」のものが第２のクラス品詞ｎ－ｇｒａｍモデルを表している。

【００６５】

【数８】

$\mathcal{M}^N_{\mathrm{class1}},\mathcal{M}^N_{\mathrm{class2}}$: **クラス品詞 *N*-gram モデル**

$$P(w_it_i|w_0t_0\cdots w_{i-1}t_{i-1}\mathcal{M}^N_{\mathrm{class1}}) \equiv P(w_i|t_i)P(t_i|c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1}) \tag{21}$$

$$P(w_it_i|w_0t_0\cdots w_{i-1}t_{i-1}\mathcal{M}^N_{\mathrm{class2}}) \equiv P(w_it_i|c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1}) \tag{22}$$

記憶長の長さＮ－１の第１のクラス品詞ｎ－ｇｒａｍモデルは、（２１）式で定義され、記憶長の長さＮ－１の第２のクラス品詞ｎ－ｇｒａｍモデルは、（２２）式で定義される。

【００６６】

記憶長の長さＮ－１の第１のクラス品詞ｎ－ｇｒａｍモデルは、品詞タグt_iをとる中でその単語ω_iが出現する条件付き確率Ｐ（$\omega_i | t_i$）と、直前Ｎ－１個の単語に係るクラス・品詞タグ列$c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1}$の並びに続いてその単語$\omega_i$の品詞タグ$t_i$が出現する条件付き確率Ｐ（$t_i | c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1}$）との積で定義される。

【００６７】

記憶長の長さＮ－１の第２のクラス品詞ｎ－ｇｒａｍモデルは、直前Ｎ－１個のクラス・品詞タグ列$c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1}$の並びに続いて、単語$\omega_i$とその品詞タグ$t_i$との組み合わせ$\omega_i t_i$が出現する条件付き確率Ｐ（$\omega_i t_i | \omega_{i-N+1}t_{i-N+1}\cdots\omega_{i-1}t_{i-1}$）で定義される。

【００６８】

このようなクラスを利用して単語の出現確率を予測することにより、品詞や語彙化した品詞とは異なる情報も用いて、仮説の生成確率を計算することが可能となっている。また、クラスを用いた形態素解析方法は既に知られているが、当該形態素解析装置５００は、上述のように、クラス品詞ｎ－ｇｒａｍ以外の確率モデルと確率的に重み付けをして結合して用いるため、クラスを用いたことによる精度の低下等の副作用が起りにくい。

【００６９】

以上のように、確率モデルにより、各仮説に対する生成確率の計算を行った後、最適解の探索を行い（６０５）、結果を出力する（６０６）。

【００７０】

図１０は、上述の生成確率計算部５１３において使用する確率モデル及び確率モデルの重みを、あらかじめ用意された品詞タグ付きコーパス及び品詞タグ無しコーパスを用いて求める処理を示すフローチャートである。

【００７１】

まず、クラス学習部５４１により、品詞タグ付きコーパス格納部５３１に格納されている品詞タグ付きコーパス及び品詞タグ無しコーパス格納部５３４に格納

されている品詞タグ無しコーパスを用いて、クラスタリングのためのパラメータを学習し、クラスタリングパラメータ格納部５４２へ格納する（７０１）。

【００７２】

但し、ここでのクラスタリングは、コーパス中の単語情報のみを用いて、その単語にクラスを与えるものとする。そのため、クラスタリングのパラメータの学習には、作成するのが困難な品詞タグ付きコーパスだけでなく容易に入手可能な品詞タグ無しコーパスを用いることができる。このようなクラスタリングを行う方法の一つとして、隠れマルコフモデルを用いることができ、この場合、Ｂａｕｍ－Ｗｅｌｃｈアルゴリズムによりパラメータの学習を行うことができる。隠れマルコフモデルの学習及びクラスの付与については、例えば、『Ｌ．Ｒａｂｉｎｅｒ，Ｂ－Ｈ．Ｊｕａｎｇ著、古井監訳、「音声認識の基礎（下）」、１９９５年』等に詳しく紹介されている。

【００７３】

次に、クラスタリングパラメータ格納部５４２中のクラスタリング用パラメータを用いて、クラス付与部５４３は、品詞タグ付きコーパス格納部５３１に格納された品詞タグ付きコーパスを入力し、各単語のクラスタリングを行い、クラスを付与し、そのクラスの付与された品詞タグ付きコーパスを品詞タグ・クラス付きコーパス格納部５３５へ格納する（７０２）。次に、確率推定部５３２により、確率モデルのパラメータを学習する（７０３）。

【００７４】

ここで、クラス品詞ｎ－ｇｒａｍモデル以外の各確率モデルに対するパラメータは、第１の実施形態の場合と同様に学習する。単語列、品詞タグ列、クラス・品詞タグ列などの系列をＸとし、その系列Ｘが品詞タグ・クラス付きコーパス格納部５３５に格納されたコーパス中に出現した回数をｆ（Ｘ）で表すと、クラス品詞ｎ－ｇｒａｍモデルに対するパラメータは、（２３）式～（２５）式のように表される。

【００７５】

【数９】

$\mathcal{M}^N_{class1}, \mathcal{M}^N_{class2}$: クラス品詞 N-gram モデル

$$P(w_i|t_i) = \frac{f(t_i w_i)}{f(t_i)} \quad (23)$$

$$P(t_i|c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1}) = \frac{f(c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1}t_i)}{f(c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1})} \quad (24)$$

$$P(w_i t_i|c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1}) = \frac{f(c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1}w_i t_i)}{f(c_{i-N+1}t_{i-N+1}\cdots c_{i-1}t_{i-1})} \quad (25)$$

記憶長の長さＮ－１の第１及び第２のクラス品詞ｎ－ｇｒａｍモデルは、上述したように、（２１）及び（２２）式で表されるので、（２１）式及び（２２）式の右辺の各要素Ｐ（ω ｉ｜ｔ ｉ）、Ｐ（ｔ ｉ｜ｃ ｉ－Ｎ＋１ｔ ｉ－Ｎ＋１…ｃ ｉ－１ｔ ｉ－１）及びＰ（ω ｉｔ ｉ｜ω ｉ－Ｎ＋１ｔ ｉ－Ｎ＋１…ω ｉ－１ｔ ｉ－１）を、（２３）式～（２５）式に従ってパラメータとして得る。

【００７６】

各確率モデルでのパラメータを確率モデル格納部５２２へ格納した後には、重み計算部５３３において重みの計算を行い、その結果を重み格納部５２３へ格納する（７０４）。

【００７７】

重みの計算については、図１１のフローチャートに示す手順で行う。第２の実施形態の重みの計算も、品詞タグ付きコーパス格納部１３１に格納されている品詞タグ付きコーパスの代わりに品詞タグ・クラス付きコーパス格納部５３５に格納されている品詞タグ・クラス付きコーパスを利用する点、品詞ｎ－ｇｒａｍ、語彙化品詞ｎ－ｇｒａｍ及び階層化品詞ｎ－ｇｒａｍに加えて、クラス品詞ｎ－ｇｒａｍを確率モデルとして用いる点を除けば、第１の実施形態の重み計算の処理（図４参照）と同様であるので、その処理の詳細説明は省略する。

【００７８】

（Ｂ－３）第２の実施形態の効果

上記第２の実施形態によれば、形態素辞書を利用して得た複数の形態素解析結果（仮説）から最尤のものを決定する際に、クラスタリングにより付与したクラス情報をも用いるようにしたので、品詞よりは細かく、語彙化した品詞よりは抽

象化された情報を利用でき、より頑健で高精度な解析を行うことができる。また、品詞タグ無しデータを利用してクラスタリングの精度を高めているので、形態素解析結果の精度も高まっている。

【００７９】

（Ｃ）他の実施形態

上記第１の実施形態では、仮説の生成確率を、品詞ｎ－ｇｒａｍ確率モデル、語彙化品詞ｎ－ｇｒａｍ確率モデル及び階層化品詞ｎ－ｇｒａｍ確率モデルを利用して求めるものを示し、第２の実施形態では、仮説の生成確率を、品詞ｎ－ｇｒａｍ確率モデル、語彙化品詞ｎ－ｇｒａｍ確率モデル、階層化品詞ｎ－ｇｒａｍ確率モデル及びクラス品詞ｎ－ｇｒａｍ確率モデルを利用して求めるものを示したが、本発明は、適用する複数種類の確率モデルの中に階層化品詞ｎ－ｇｒａｍ確率モデルが含まれていれば、複数種類の確率モデルの組み合わせは、上記実施形態のものに限定されない。

【００８０】

また、仮説生成部１１２、５１２による仮説（形態素解析結果候補）の生成方法は、形態素辞書を利用した一般的な形態素解析方法に限定されず、文字に関するｎ－ｇｒａｍを利用した形態素解析方法など、他の形態素解析方法を利用するようにしても良い。

【００８１】

さらに、上記各実施形態では、最尤の仮説である形態素解析結果を出力するものを示したが、得られた形態素解析結果を、機械翻訳部などの自然言語処理部に直ちに与えるようにしても良い。

【００８２】

さらにまた、上記各実施形態では、モデル学習部やクラスタリング部を備えるものを示したが、モデル学習部やクラスタリング部を備えないで、解析部とモデル格納部とで形態素解析装置を構成するようにしても良い。この場合、モデル格納部への情報は、予めモデル学習部やクラスタリング部で形成されたものである。また、第２の実施形態でクラスタリング部などを省略した場合には、モデル格納部にクラス付与機能を持たせることを要する。

【００８３】

また、各種の処理に供するコーパスは、通信処理により、ネットワークなどから取り込むようなものであっても良い。

【００８４】

本発明が適用可能な言語は、上記実施形態のような日本語には限定されないことは勿論である。

【００８５】

【発明の効果】

以上のように、本発明によれば、複数の正解候補の中から最適な解を高い精度で選択し得る形態素解析装置、形態素解析方法及び形態素解析プログラムを提供できる。

【図面の簡単な説明】

【図１】

第１の実施形態の形態素解析装置の機能的構成を示すブロック図である。

【図２】

第１の実施形態の形態素解析装置の解析時動作を示すフローチャートである。

【図３】

第１の実施形態の形態素解析装置のモデル学習動作を示すフローチャートである。

【図４】

図３の重みの計算処理の詳細を示すフローチャートである。

【図５】

第１の実施形態のモデルパラメータの例を示す説明図（その１）である。

【図６】

第１の実施形態のモデルパラメータの例を示す説明図（その２）である。

【図７】

第１の実施形態のモデルパラメータの例を示す説明図（その３）である。

【図８】

第２の実施形態の形態素解析装置の機能的構成を示すブロック図である。

【図９】

第２の実施形態の形態素解析装置の解析時動作を示すフローチャートである。

【図１０】

第２の実施形態の形態素解析装置のモデル学習動作を示すフローチャートである。

【図１１】

図１０の重みの計算処理の詳細を示すフローチャートである。

【符号の説明】

１００、５００…形態素解析装置、

１１０、５１０…解析部、

１１２、５１２…仮説生成部、１１３、５１３…生成確率計算部、

１１４、５１４…解探索部、

１２０、５２０…モデル格納部、

１２１、５２１…形態素辞書格納部、１２２、５２２…確率モデル格納部、

１２３、５２３…重み格納部、

１３０、５３０…モデル学習部、

１３１、５３１…品詞タグ付きコーパス格納部、

１３２、５３２…確率推定部、１３３、５３３…重み計算部、

５３４…品詞タグ無しコーパス格納部、

５３５…品詞タグ・クラス付きコーパス格納部、

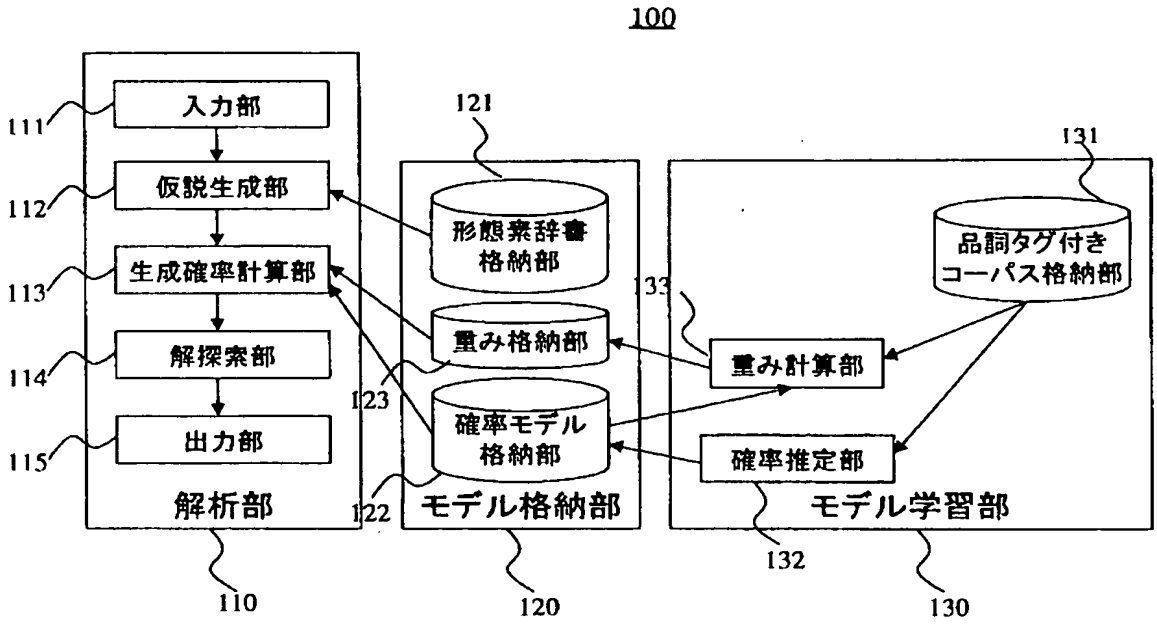
５４０…クラスタリング部、

５４１…クラス学習部、５４２…クラスタリングパラメータ格納部、

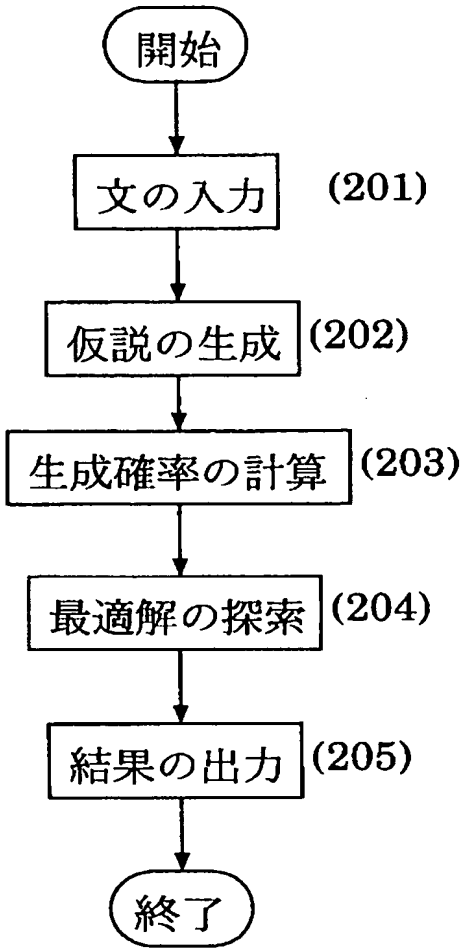
５４３…クラス付与部。

【書類名】　図面

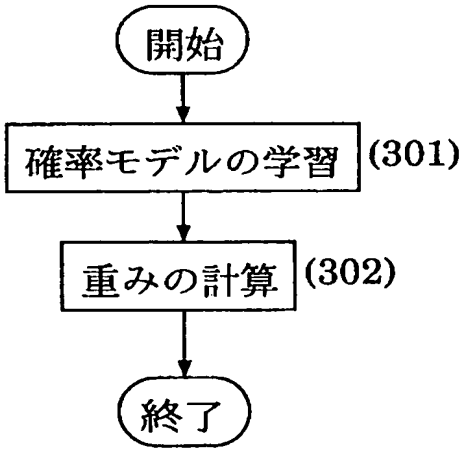
【図１】


100
入力部
111
仮説生成部
112
生成確率計算部
113
解探索部
114
出力部
115
解析部
110
121
形態素辞書格納部
重み格納部
123
確率モデル格納部
122
モデル格納部
120
131
品詞タグ付きコーパス格納部
133
重み計算部
確率推定部
132
モデル学習部
130

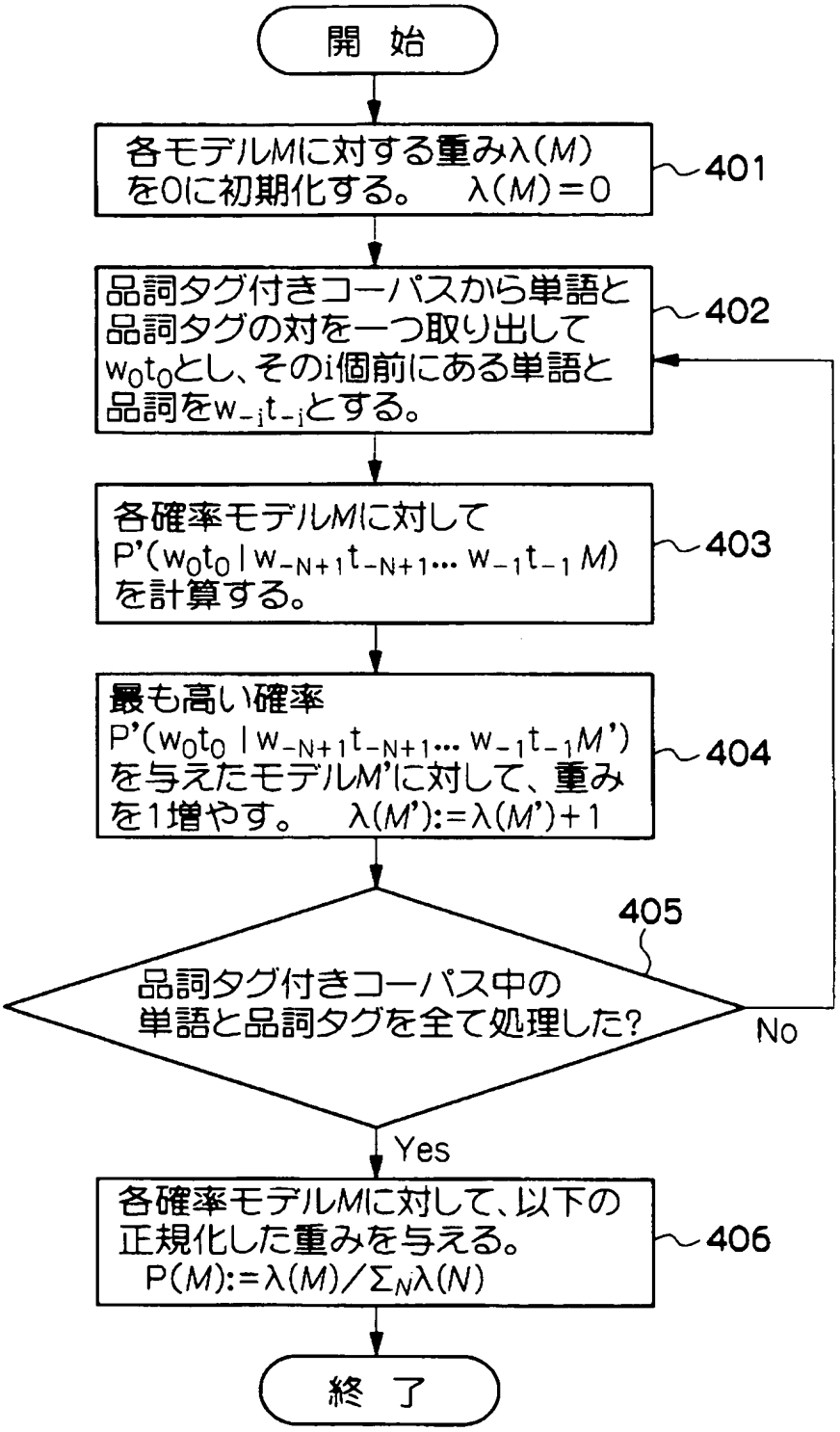
【図２】


開始
文の入力
(201)
仮説の生成
(202)
生成確率の計算
(203)
最適解の探索
(204)
結果の出力
(205)
終了

【図３】


開始
確率モデルの学習
(301)
重みの計算
(302)
終了

【図４】


開　始
各モデルMに対する重みλ(M)を0に初期化する。　λ(M)＝0
401
品詞タグ付きコーパスから単語と品詞タグの対を一つ取り出してw_0t_0とし、そのi個前にある単語と品詞を$w_{-i}t_{-i}$とする。
402
各確率モデルMに対して$P'(w_0t_0 | w_{-N+1}t_{-N+1}\cdots w_{-1}t_{-1}M)$を計算する。
403
最も高い確率$P'(w_0t_0 | w_{-N+1}t_{-N+1}\cdots w_{-1}t_{-1}M')$を与えたモデルM'に対して、重みを1増やす。　λ(M'):=λ(M')+1
404
405
品詞タグ付きコーパス中の単語と品詞タグを全て処理した?
No
Yes
各確率モデルMに対して、以下の正規化した重みを与える。
$P(M):=\lambda(M)/\Sigma_N\lambda(N)$
406
終　了

【図５】

P(政権 | 名詞-普通名詞)=0.004343
P(裁判官 | 名詞-普通名詞)=0.000328
P(中国語 | 名詞-普通名詞)=0.000036
P(コート | 名詞-普通名詞)=0.000018
P(倫理 | 名詞-普通名詞)=0.000182
P(騒然と | 形容詞:基本連用形)=0.001216
P(悠長な | 形容詞:ダ列基本連体形)=0.000515
P(手渡さ | 動詞:未然形)=0.000319
P(損なわ | 動詞:未然形)=0.000559
P(結集し | 動詞:基本連用形)=0.000178
・
・
・

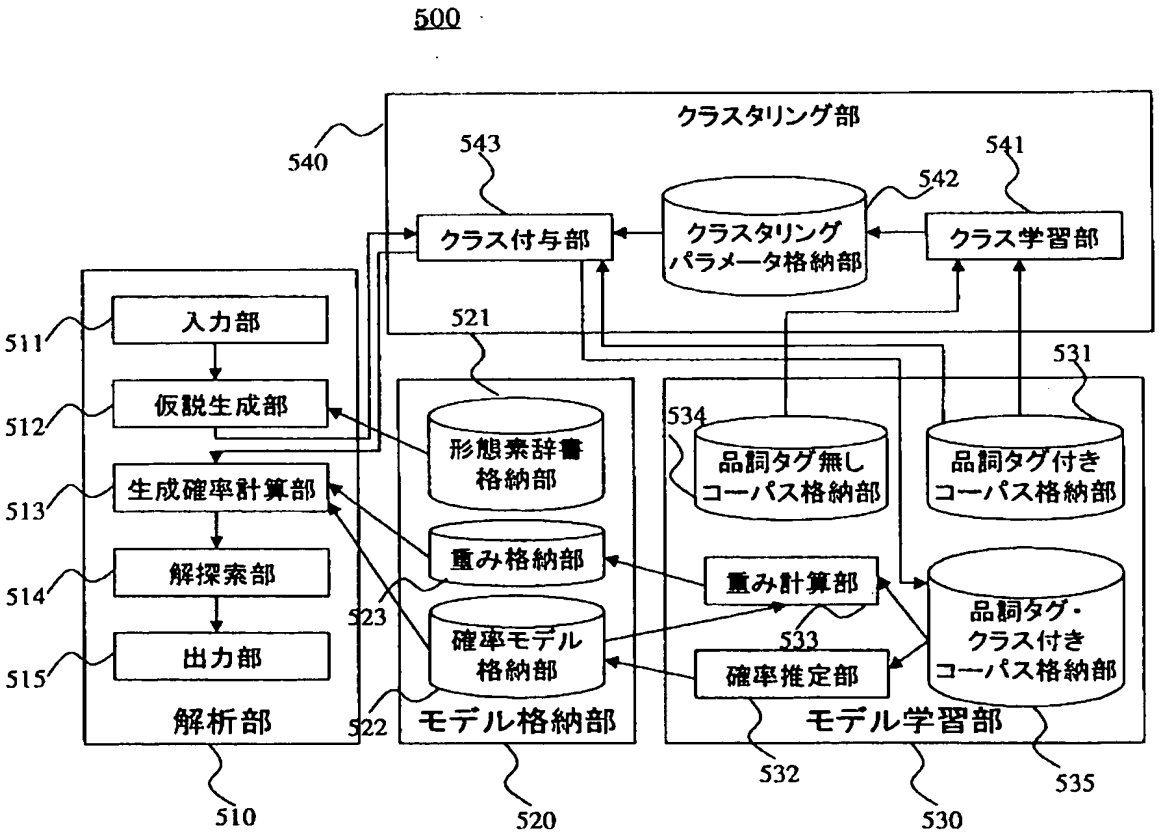
【図６】

P(名詞-普通名詞 | 名詞-数詞)=0.075674
P(副詞 | 副詞)=0.012787
P(特殊-句点 | 助詞-終助詞)=0.734557
P(名詞-固有名詞 | 副詞)=0.000314
P(動詞:未然形 | 副詞)=0.025575
P(形容詞:基本形 | 副詞)=0.033498
P(名詞-サ変名詞 | 副詞)=0.129285
P(動詞:基本形 | 副詞)=0.034832
P(特殊-句点 | 形容詞:基本推量形)=0.857143
P(特殊-読点 | 動詞:基本連用形)=0.543240
・
・
・

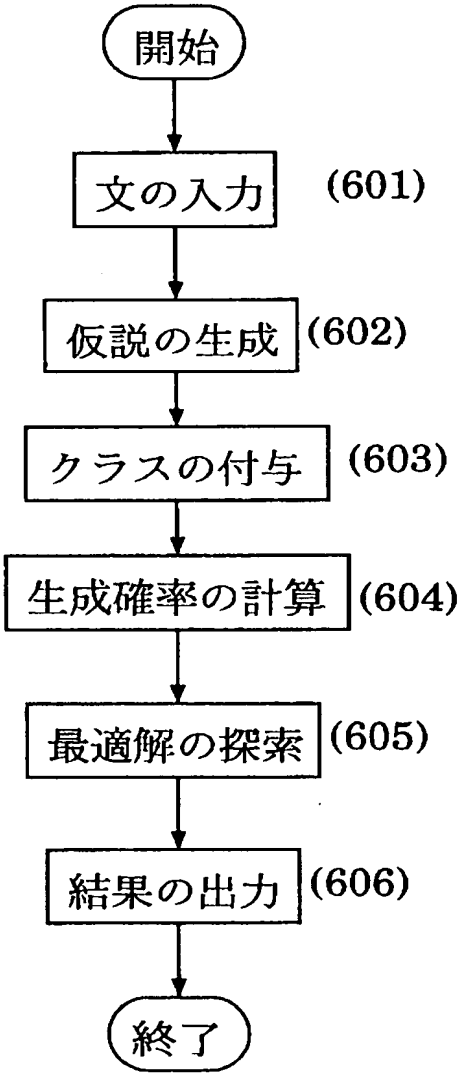
【図７】

P(。__特殊-句点 | つかんだ__動詞:夕形)=0.562500
P(に__助詞-格助詞 | 代わり__名詞-普通名詞)=0.782609
P(多方__名詞-人名 | は__助詞-副助詞)=0.000032
P(した__動詞:夕形 | 録音__名詞-サ変名詞)=0.285714
P(が__助詞-格助詞 | 漁船__名詞-普通名詞)=0.166667
P(不必要だ__形容詞:基本形 | 全く__副詞)=0.010638
P(市場__名詞-普通名詞 | 海外__名詞-普通名詞)=0.021858
P(教室__名詞-普通名詞 | で__助詞-格助詞)=0.000154
P(は__助詞-副助詞 | まで__助詞-接続助詞)=0.060000
P(を__助詞-格助詞 | 實務__名詞-普通名詞)=0.461538
・
・
・

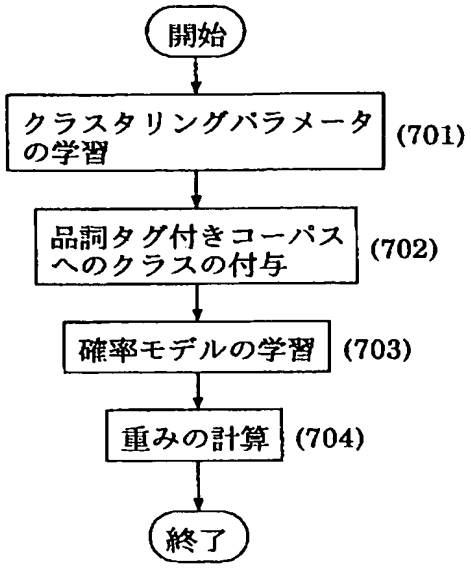
【図８】



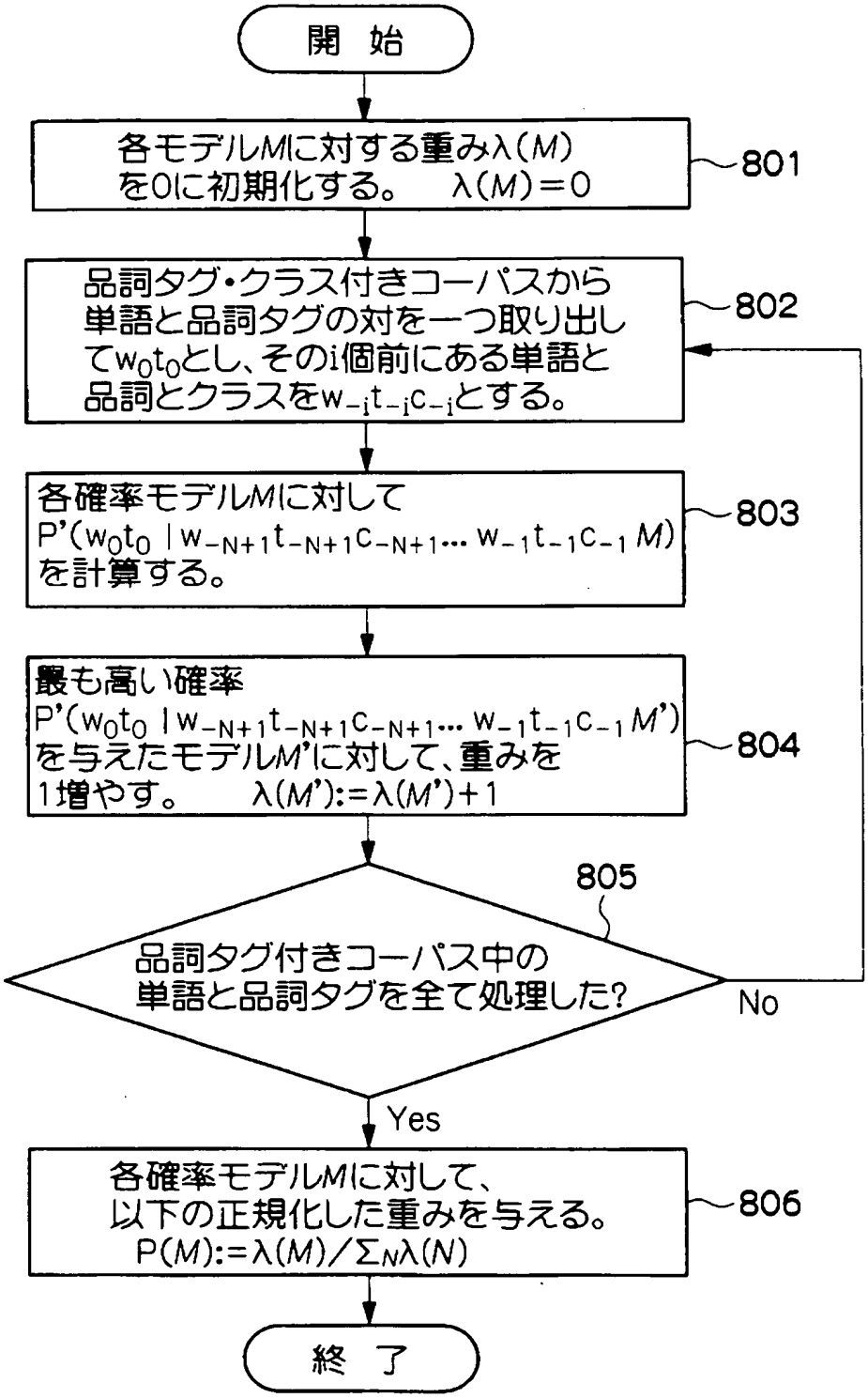
【図９】


開始
文の入力
(601)
仮説の生成
(602)
クラスの付与
(603)
生成確率の計算
(604)
最適解の探索
(605)
結果の出力
(606)
終了

【図１０】


開始
クラスタリングパラメータの学習
(701)
品詞タグ付きコーパスへのクラスの付与
(702)
確率モデルの学習
(703)
重みの計算
(704)
終了

【図１１】


開　始
各モデルMに対する重みλ(M)
を0に初期化する。　λ(M)＝0
801
品詞タグ・クラス付きコーパスから
単語と品詞タグの対を一つ取り出し
てw0t0とし、そのi個前にある単語と
品詞とクラスをw-it-ic-iとする。
802
各確率モデルMに対して
P'(w0t0 | w-N+1t-N+1c-N+1… w-1t-1c-1 M)
を計算する。
803
最も高い確率
P'(w0t0 | w-N+1t-N+1c-N+1… w-1t-1c-1 M')
を与えたモデルM'に対して、重みを
1増やす。　λ(M'):=λ(M')+1
804
805
品詞タグ付きコーパス中の
単語と品詞タグを全て処理した?
No
Yes
各確率モデルMに対して、
以下の正規化した重みを与える。
P(M):=λ(M)／ΣNλ(N)
806
終　了

【書類名】　　要約書

【要約】

【課題】　複数の正解候補の中から最適な解を高い精度で選択し得る形態素解析装置、形態素解析方法及び形態素解析プログラムを提供する。

【解決手段】　本発明では、文を形態素解析して得た形態素解析結果の候補である各仮説に対し、品詞ｎ－ｇｒａｍ確率モデル、語彙化した品詞に係る語彙化品詞ｎ－ｇｒａｍ確率モデル、及び、品詞を品詞本体と品詞の活用形とに階層化して求められている階層化品詞ｎ－ｇｒａｍ確率モデルを重み付け結合して、生成確率を求め、各仮説の生成確率に基づいて、解（最終的な形態素解析結果）を探索する。

【選択図】　　図１

特願２００３－１５４６２５

出　願　人　履　歴　情　報

識別番号　　［０００００２９５］

１．変更年月日　　１９９０年　８月２２日
［変更理由］　　新規登録
住　所　　東京都港区虎ノ門１丁目７番１２号
氏　名　　沖電気工業株式会社